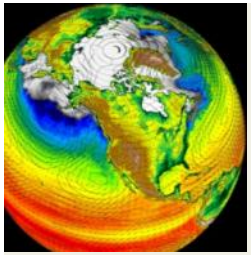
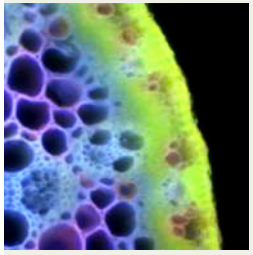
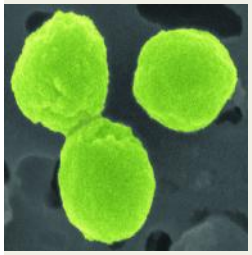


Data Management Considerations and Challenges

(Annual ESGF Face-to-Face Meeting)

Justin Jay Hnilo

Program Manager - Data Management

CESD/BER/SC

December 6-9 , 2016

When we think about data, we need to think about requirements that should be gathered?

- **Functional:** What should the end resulting product do.
- **Data requirements:** Capture the type, volatility, size/amount, persistence, accuracy and the amounts of the required data.
- **Environmental requirements:** a) content of use; b) Social environment (e.g., collaboration and coordination); c) how good is user support likely to be d) what technologies will it run on
- **User Requirements:** Capture the characteristics of the intended user groups.
- **Usability Requirements:** Usability goals associated measures for a particular products used in CESD.

CESD supports diverse research programs.

Example of Some Key CESD Data Resources

Data generating projects and holdings	Description
ACME	Designing an operational test bed for advanced Earth system model development; has among the most varied data management needs including management of diagnostics and regridded output and provenance tracking.
AmeriFlux	AmeriFlux is a network of ~150 past and present terrestrial sites in Central, North, and South America making continuous measurements of carbon dioxide, radiation, and water vapor fluxes along with continuous meteorological measurements. Station data with various data versions and product levels. Station data at multiple heights and depths. Measurement frequencies vary from continuous eddy-covariance measurements to infrequent biological samplings.
ARM	The ARM Data Archive collects and distributes over 4000 ARM observational and PI data products. The ARM data management includes data collection and preparation, data quality, data and metadata dissemination services, metrics collection and reporting, data processing as-a-service. In addition, ARM has a well defined metadata architecture which is used in data discovery and access.
CAPT	Gridded forcing data are high-resolution analyses. Outputs from models are either limited spatial domain regions or station data-like output.
CDIAC	CDIAC's data collection covers numerous disciplines (chemical oceanography, meteorology, climatology, atmospheric chemistry), time periods, and geographic representations (global to microenvironments). The collection includes original data and derived, value-added data products.
CMIP (PCMDI)	Gridded GCM data. Many temporal resolutions.
FACE	Station data at multiple heights and depths. Measurement frequencies vary from continuous measurements to infrequent biological samplings.
ILAMB	Gridded data as well as station data, high temporal frequency.
NGEE Arctic Data Archive	Measurements are made and samples are collected at multiple heights and depths. Measurement frequencies vary from continuous measurements to infrequent biological samplings. Station data and merged products.
Obs4MIPs and 70 other model intercomparison projects (MIPs)	Gridded data of comparable form to CMIP efforts.
SPRUCE Data Archive	Measurements are made and samples are collected at multiple heights and depths. Measurement frequencies vary from continuous measurements to infrequent biological samplings. Station data and merged products.

The Data Challenge

- **Isolated** - data resides within programs, facilities and ongoing community research projects
- **Specialized** - Need for a unified capability to cross talk data outside of individual research domains
- **Exponential Growth** – Continual increases in the volume, acquisition rate, variety, and complexity of scientific data
- **Common Language - Data Management & Standards**
  - Cross-disciplinary science working across techniques, integrating simulation, and experimental/observational results complicates data management, analysis, and visualization
  - Metadata standards are varied or non-existent
- **Overhead Costs** - Too much time is spent rewriting data to a usable form.

These all act to complicate the accessibility, availability and usefulness of high quality research data to address multi-disciplinary problems.

Capability Components

Data Integration

1) Integrating complex data generating systems
2) High throughput networks
3) Data collection and management

Computational Environment

4) Data analytics
5) Simplified and modern user interfaces
6) Decision control and knowledge discovery

Data Integration Requirements

Data Strategy

- Create integrated Architecture
  - BER/CESD modeling efforts, field observations, and laboratory experimental results
- Allow users to find what the need easily with search and modern user interfaces
- Develop libraries to allow cross talk between data repositories
- Develop consistent metadata

Component requirements provided by data experts at National Laboratories and DOE Program Managers.

Computational Environment Requirements

Computational Data Analytics Strategy

- Leverage existing and future DOE leadership class facilities
- Implement an analysis platform
- Develop visualization/intercomparison tools
- Provide Provenance, automation, and human-computer interaction

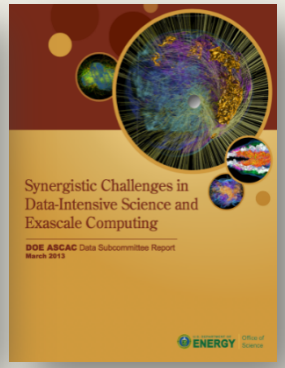
NEED: Input for this component will require extensive executive committee and community involvement.

Virtual Laboratory Concept
(Integrated Data Environment)

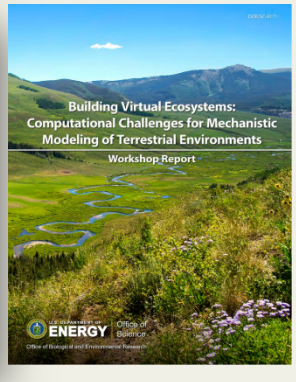
Inputs to CESD/BER: BERAC Virtual Lab Report, ASCAC data workshop, ESS Cyberinfrastructure Working Group workshop, ESGF conference.

Community involvement and outreach define our next steps

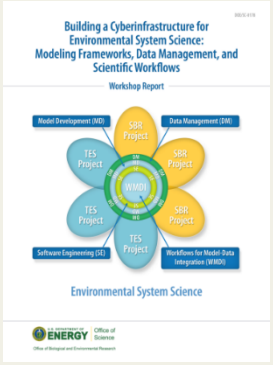
2013



2013

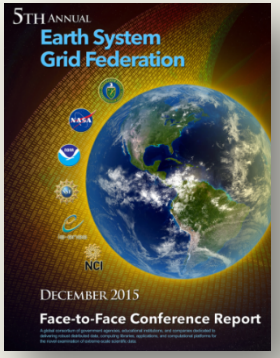


2014



2015

2015



2015

SC Statement on Digital Data Management
(http://science.energy.gov/funding-opportunities/digital-data-management/)

-All proposals submitted to the Office of Science for research funding must include a DMP.
-Involves all stages of the digital data life cycle: capture, analysis, sharing, and preservation.

A Virtual Laboratory Vision

Global Climate Modeling

Regional Modeling

ARM

ESS

EMSL

KBase.

DOE BER Data Infrastructure

Transfer

External Data Centers

API's and Services

Data Center 1.0

Data Center 1.1

Data Center 1.2

Data Center 1.n

Public Access

Provenance Workflow

Visualization Framework

Security

Data & metadata services

Data Center

Private Access

Metrics

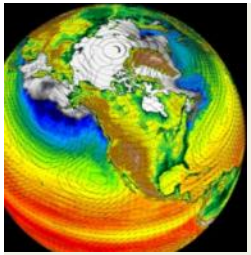
Data Quality

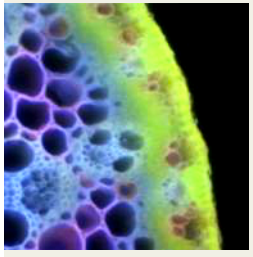
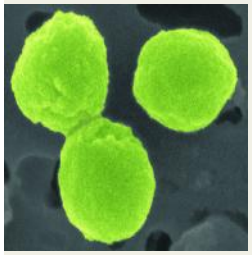
Simulation

Integrated Data Environment

An Integrated Cyber-infrastructure leveraging core Office of Science resources to enable discovery, analytics, simulation, and knowledge innovation

Data Center & Interoperable Services



Thank you!

Contact:
justin.hnilo@science.doe.gov